University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 14: Decision Trees, Practical Exploration of Classifiers

Exercise 14-1 Decision trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1-2$ years, $2-7$ years, >7 years)
- Gender (male, female)
- Residential area (urban, rural)

For your analysis you have the following manually classified training examples:

Person	Time since license	Gender	Area	Risk class
1	$1-2$	m	urban	low
2	$2-7$	m	rural	high
3	>7	f	rural	low
4	$1-2$	f	rural	high
5	>7	m	rural	high
6	$1-2$	m	rural	high
7	$2-7$	f	urban	low
8	$2-7$	m	urban	low

(a) Construct a decision tree based on this training dataset. Use information gain for selecting the split attributes. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class, you do not need to apply a pruning algorithm.

(b) Apply the decision tree to the following drivers:
Person A: 1-2, f, rural
Person B: 2-7, m , urban
Person C: 1-2, f, urban

Exercise 14-2 Decision trees, naïve Bayes, and k-nn classification – Practical

Bring your laptop for this interactive lab session in the exercise class.

(a) Work with some toolbox for classification (e.g., R, Python, WEKA) to study the impact of different settings on the behavior of decision trees, the naïve Bayes classifier, and the k nearest neighbor classifier on some dataset (e.g., Iris).

(b) How does the behavior of the k nearest neighbor classifier change with the choice of k?

(c) What is the impact of parameter choices on the quality of decision trees?

(d) How does the behavior of the three classifiers change with the amount of training data (e.g., choice of training-test-splits)?